1 – Introdução

Este Projeto de Formação Complementar teve por objetivo introduzir o aluno à linguagem e programação Python e ferramentas utilizadas para a análise de dados utilizando a linguagem.

O estudo se deu inicialmente com uma introdução à sintaxe de Linguagem Python: funcionamento das variáveis e seus tipos, estruturas de dados *built-in;* funcionamento de operadores condicionais, laços de repetição e funções; e funcionamento de Programação Orientada à Objetos. Além de tratamento de arquivos e pacotes de Python.

Após o estudo geral da linguagem, debruçou-se sobre o estudo de bibliotecas voltadas para Análise de dados, tais como: NumPy, Pandas, Matplolib, SciPy, Seaborn.

Para a montagem deste relatório, foi feito um estudo de caso e análise exploratória de um *Dataset* retirado do site *Kaggle*. Este site é uma base de dados que reúne milhares de *datasets* dos mais variados tipos para qualquer pessoa fazer um estudo sobre eles.

2 – Materiais e Métodos

O *dataset* reúne dados sobre a data da coleta, o consumo de cerveja no dia e as temperaturas aferias no dia na cidade de São Paulo, durante todo o ano de 2015. Os dados foram coletados em um bairro universitário, onde grupos de estudantes entre 18 e 28 anos fazem festas.

Sobre o *dataset* temos 7 diferentes atributos: a data da coleta, as temperaturas máxima, média e mínima da cidade de São Paulo no dia, se a data da coleta era um final de semana ou não, precipitação no dia, e o consumo de cerveja no dia em milhares de litros.

O método de análise dos dados utilizado foi a análise exploratória. A análise exploratória, segundo a IMB (International Business Machines Corporation), é utilizada por cientistas de dados para investigar conjuntos de dados e resumir suas principais características.

O método de análise exploratória consiste em transformar os dados coletados em informação estruturada que pode ser visualizada. Isso permite que os analistas de dados tenham novas ideias e visões sobre os dados coletados para apoiar a tomada de decisões.

3 – Apresentação dos Resultados de Análise

Considerando os dados organizados no *dataset* inicialmente foi necessário mudar o tipo de dado de algumas colunas de dados *string* para dados inteiros ou de ponto flutuante, ou até mesmo booleanos. Para se ter consistência nos dados foi verificado de não existiam dados nulos

ou vazios e duplicatas, se fossem encontrados seriam removidos. A Figura 1 mostra a distribuição do consumo de cerveja em dezenas de milhas de litros ao longo do ano de 2015:

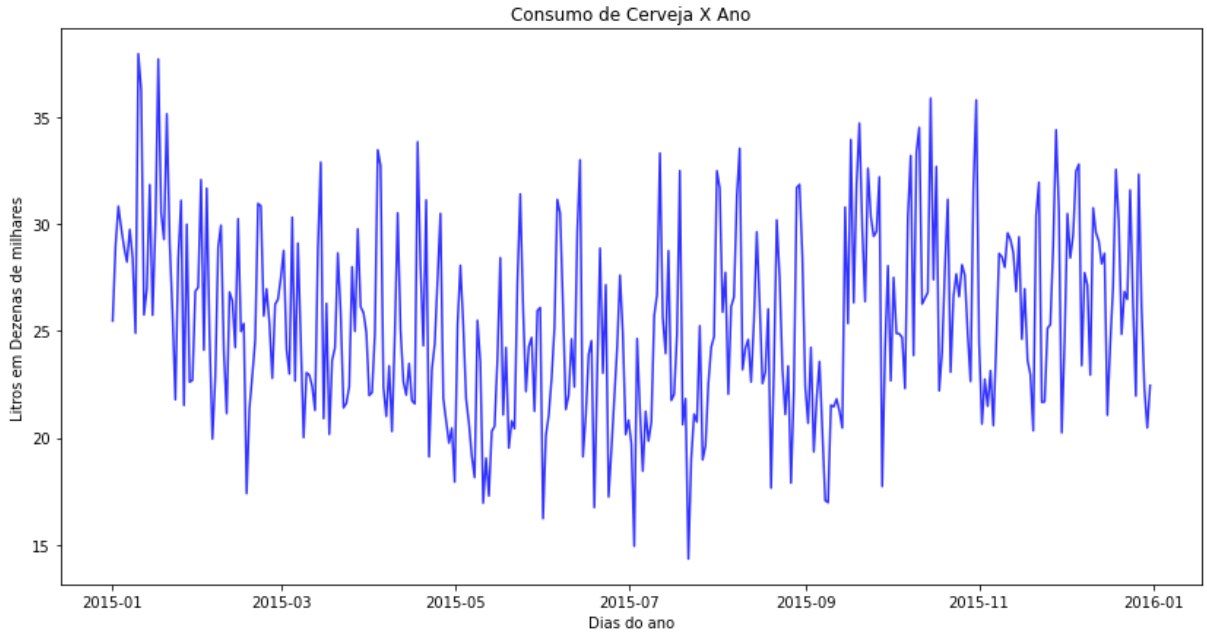

Figura 01 – Consumo de Cerveja ao longo do ano de 2015.

Assim, foi montado um *boxplot* para verificar a mesma distribuição do consumo de cerveja ao longo do ano, porém separado por meses, como mostra a Figura 2:

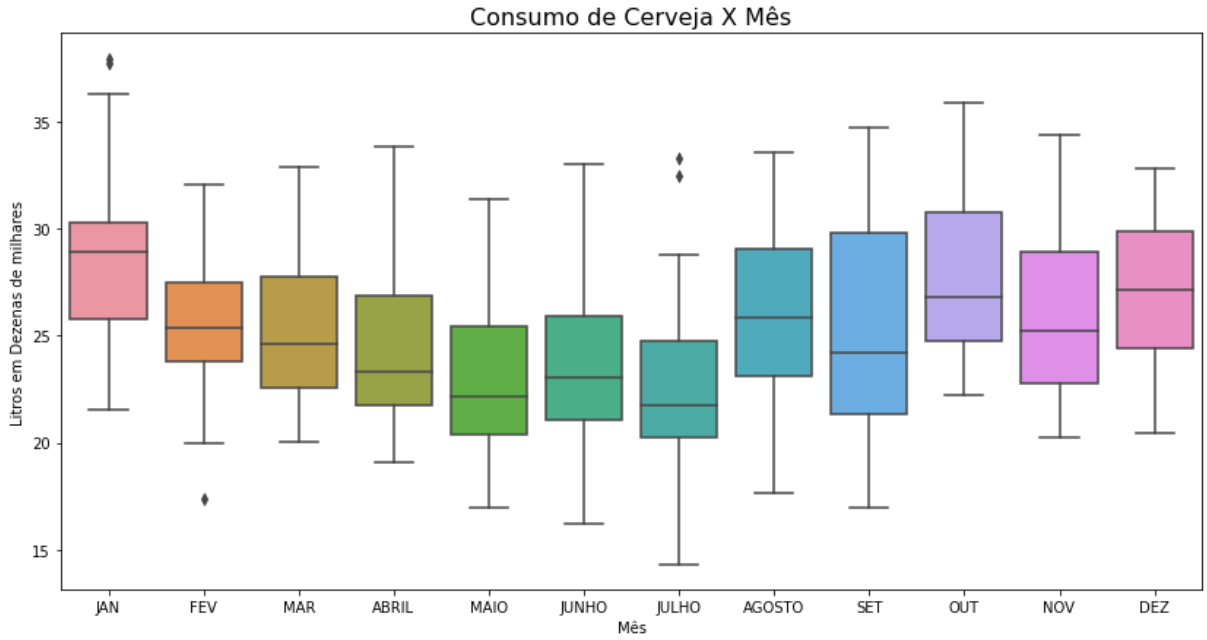

Figura 2 – Consumo de Cerveja ao longo do ano de 2015 separado por mês.

A partir da montagem desses dois gráficos fica evidente que nos meses em que é Outono e Inverno, ou seja, do final do mês de março até meados de setembro, existem consumos menores que se comparados aos outros meses do ano, em que é Primavera e Verão.

Desta maneira, podemos tecer uma comparação do consumo de cerveja em faixas de temperatura aferidas na cidade de São Paulo, para cada uma das medições máxima, média e mínima. A Figura 3 mostra a divisão das faixas de temperatura mínima pelos quartis de temperaturas em: menor ou igual que 15ºC, maior que 15ºC e menor ou igual que 18ºC, maior que 18ºC e menor ou igual a 10ºC, e maior que 20ºC; e mostra também a divisão entre menor ou igual que 18ºC e maior que 18ºC:

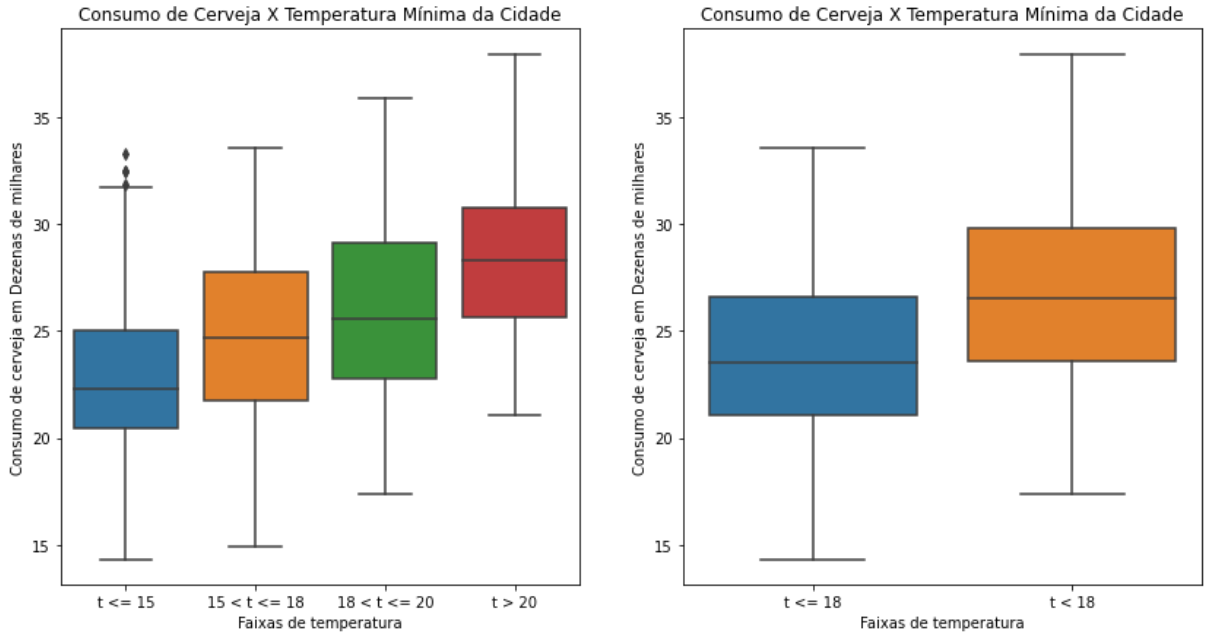

Figura 3 – Distribuição do consumo de cerveja por faixas de temperatura mínima.

Agora, fazendo a divisão das temperaturas máximas também em quartis: menor ou igual a 23ºC, maior que 23ºC e menor ou igual a 16ºC, maior que 26ºC e menor ou igual a 29ºC, e maior que 29ºC; e a divisão entre menor ou igual a 26ºC e maior que 26ºC. Montou-se um novo gráfico de *boxplot* com essas divisões, como mostra a Figura 4:

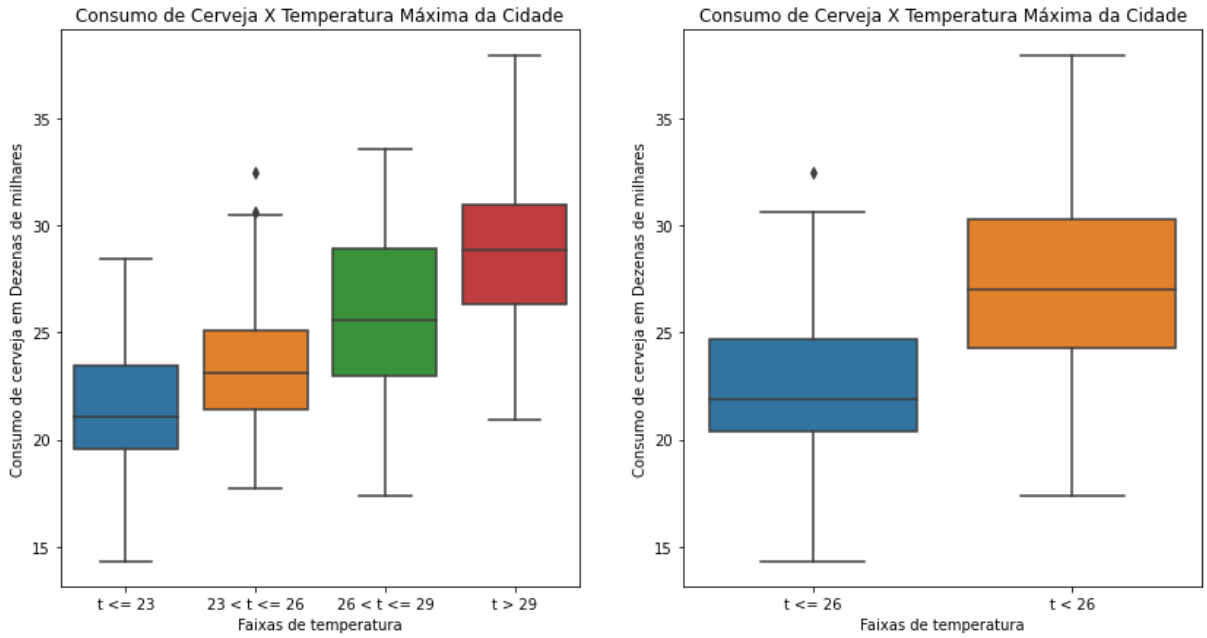

Figura 4 – Distribuição do consumo de cerveja por faixas de temperatura máxima.

Com isso, podemos observar que em dias que são mais quentes o consumo de cerveja tende a aumentar bastante. Porém, existe uma relação intrínseca entre as temperaturas mínimas da cidade e as temperaturas máximas, já que os gráficos são muito próximos um do outro.

Isso se deve ao fato de que, a cidade de São Paulo possui o clima subtropical úmido, de acordo com a classificação de Köppen proposta em 1900. Este clima determina uma temperatura média de 18ºC durante o ano e amplitude térmica de aproximadamente 10ºC, o que condiz com as divisões de temperatura dos gráficos.

Outra relação possível de ser feita com o consumo de cerveja na cidade de São Paulo é com a precipitação. A Figura 5 mostra a relação qualitativa do consumo de cerveja com a na cidade de São Paulo:

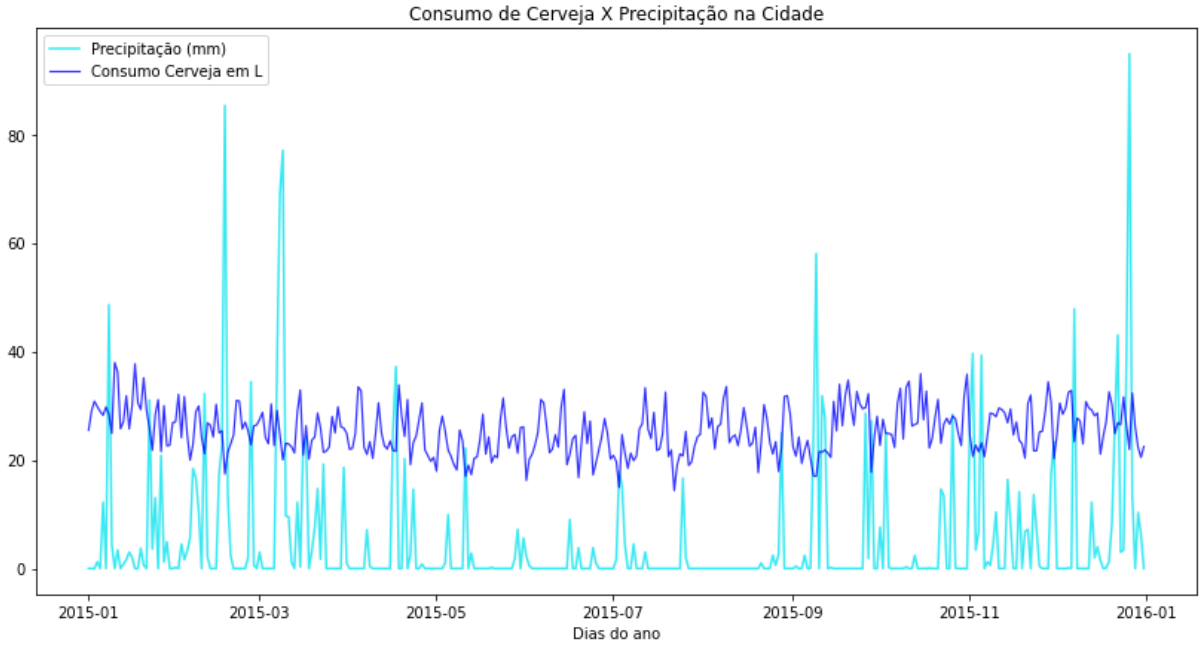

Figura 5 – Consumo de cerveja relacionado com a Precipitação

Na análise da Figura 5 fica evidente que durante dias muito chuvosos o consumo é bem menor que outros dias, sendo alguns dos menores registros feitos em dias chuvosos. Esse tipo de distribuição da precipitação também é típico do clima subtropical úmido.

4 – Conclusões

Com o estudo de caso foi possível exercitar o que foi estudado durante o Projeto de Formação Complementar e com isso foi possível entender que a linguagem de Programação Python é uma linguagem muito poderosa para realizar diversos tipos de atividade, principalmente para a análise de dados.

Pelo fato de a Linguagem de Programação Python ser de alto nível, ela possui inúmeras ferramentas e bibliotecas que facilitaram o estudo de caso, o que seria muito mais difícil com outras linguagens de baixo nível. É de suma importância no mundo atual conhecer ferramentas e formas de fazer diferentes atividades.